A Distributed Information Services Architecture to Support Biomarker Discovery in Early Detection of Cancer

Daniel Crichton[1], Sean Kelly[1], Chris Mattmann[1, 2], Qing Xiao[1], J. Steven Hughes[1], Jane Oh[1], Mark Thornquist[3], Donald Johnsey[4], Sudhir Srivastava[4], Laura Essermann[5], and William Bigbee[6]

[1]*Jet Propulsion Laboratory*
California Institute of Technology
Pasadena, CA, USA 91109
{crichton,kelly,qxiao,hughes,joh}@jpl.nasa.gov

[2]*Computer Science Department*
University of Southern California
Los Angeles, CA 90089
mattmann@usc.edu

[3]*Fred Hutchinson Cancer Research Center*
Seattle, WA 98109
thornquist@fhcrc.org

[4]*National Cancer Institute*
National Institutes of Health
Bethesda, MD 20892
srivasts@mail.nih.gov

[5]*University of California, San Francisco*
San Francisco, CA 94143
laura.esserman@ucsfmedctr.org

[6]*University of Pittsburg*
Pittsburgh, PA 15213
bigbeewl@upmc.edu

Abstract

Informatics in biomedicine is becoming more and more interconnected with distributed information services, interdisciplinary correlation, and cross-institutional collaboration. Partnering with NASA, the Early Detection Research Network, a program managed by the National Cancer Institute, has been defining and building emerging informatics architecture to support the discovery of biomarkers in their earliest stages. The architecture established by the informatics teams working with the National Cancer Institute and the Early Detection Research Network serves a blueprint for constructing a set of services focused on the capture, processing, management and distribution of information through the phases of biomarker discovery and validation. This focus of this paper will be to define the architecture, the associated services, the overall information model and framework for which these services can be integrated to form a knowledge grid for the scientific community

1. Introduction and Scientific Drivers

Scientific discovery in biomedical research increasingly depends on mining and correlating diverse data sets from various resources. Yet, the lack of an overarching model for describing the information makes it difficult to integrate highly divergent sets of information which is often locked away in computing silos. The Early Detection Research Network (EDRN) of the National Cancer Institute in the United States has established a research network led by scientists from multiple disciplines across 30 cancer centers into a collaborative research structure focused on cancer biomarker discovery. The distributed nature of the EDRN represents a modern challenge for building a bioinformatics infrastructure to capture and distribute science and ancillary data within the network. NASA's Jet Propulsion Laboratory (JPL), one of the key players of the EDRN Informatics Working Group, has been leading the ongoing effort in architecting and implementing such a biomedical data management and delivery infrastructure. The first application focused on providing a common informatics framework for accessing heterogeneous bio-specimen repositories located at participating EDRN sites across the United States. As the infrastructure has evolved, the core principles of building common services that integrate general client applications with heterogeneous, distributed data resources have not changed. Principally, the EDRN has recognized the need to build a knowledge system where bio-specimens, scientific data, study specific data, and biomarker data can be

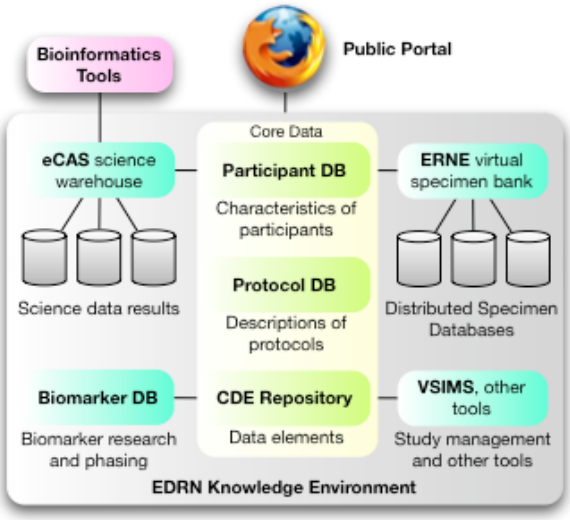
captured, accessed and shared at a national level via transparent, grid-type architecture. As a result, EDRN is focused on addressing these critical informatics goals: (1) defining a common information model for describing the EDRN information space; (2) enabling all components of the knowledge system to be distributed; (3) providing software interfaces for capture, discovery and access of data resources across the knowledge system; (4) providing a secure transfer and distribution infrastructure to meet federal regulations for data sharing; and (5) providing an integrated portal environment to allow users to seamlessly search, locate and correlate information across the distributed EDRN.

2. Architecture

One of the critical characteristics of the architecture has been leveraging common architectural patterns across very different science environments [1, 2]. Due to JPL's involvement, the architectural models and associated software services for supporting planetary science have been leveraged to develop the EDRN informatics architecture. The approach of defining the architecture focused on definition of both the information and functional portions of that EDRN knowledge system. In effect, it looked at the separation of these architectural models such that common information services developed by JPL for planetary science could be directly reused, if an abstracted model for cancer research in early detection could be developed and deployed.

As part of defining the architecture for any science-oriented data system, we identified these particular functions as having common patterns for which a set of core services could be implemented: data capture, data discovery, data access, data retrieval, data processing, and data distribution. Furthermore, each of the services that implement these functions should be distributable meaning that they should allow for distributed deployment yet work in concert with one another to allow for virtual systems to be constructed.

In addition to the functional architecture, there is also the information architecture that is critical to forming an integrated informatics platform. As part of the developing the information architecture, the EDRN has been developing a common information model for the representation of information associated with the data objects it manages within the knowledge system. EDRN developed a set of common data elements (CDEs) for describing these data and their associated attributes in order to provide a common language for communicating information across the EDRN scientific discovery lifecycle. EDRN CDEs are data elements that have been agreed upon by the EDRN investigators as critical data that must be collected by all EDRN sites; they include data elements that describe the major objects of information captured by the EDRN. The following figure demonstrates the complete knowledge environment:



In addition to defining the functional and information architectures, the EDRN worked with JPL to adopt the Object Oriented Data Technology (OODT) framework as the foundation for the EDRN informatics infrastructure. OODT provides a set of core services that implement the above functions driven by a domain model. This allows for the deployment of core services to support planetary science as well as cancer research. Each of the core component services can be deployed independently and integrated using XML-based interfaces over a distributed, grid architecture. This distributed framework, for example, makes it possible to query multiple institutional repositories concurrently, compiling the results into a unified view and making them available for analysis.

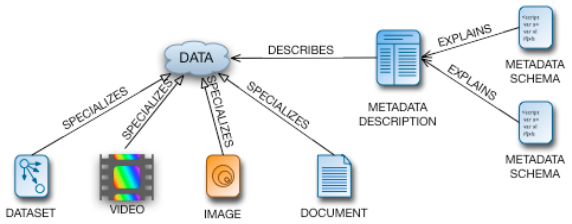
The OODT service framework is based on the software architectural notion of components. Each component has well known interfaces that enable them to be plugged together in a distributed manner. The components themselves sit on top of off-the-shelf middleware (such as SOAP services) so that they can be deployed into an enterprise topology.

The **Catalog and Archive Service** component provides the ability to catalog, process, and store information objects in a distributed environment. The **Profile Service** component provides a registry of information about managed information objects necessary to discover them. Multiple profile services can be distributed and integrated into a directed graph topology in order to crawl the registries and locate critical information objects. The **Product Service** provides a mechanism for access, retrieval, and transformation of science data products and information from remote repositories. The **Query Service** provides an interface to distributed components so that they work together.

EDRN is working towards building the knowledge system on top of these services. Application components of the knowledge system include a system for sharing biospecimens across institutions, a set of services for deploying a distributed warehousing and workflow system to capture and process scientific results from validation studies, a database for managing information about cancer biomarkers, and an integrated portal which serves as a gateway to the information stored within these application components.

3. Information Model

The domain information model is critical to describing the data objects that are captured and managed within the EDRN knowledge system. At the core of this model is identifying a standard set of metadata elements that can be used for annotating these objects. Without metadata, there is little chance of discovering data, interpreting it correctly, and reusing it in an automated fashion. As an example, consider the figure below which identifies how certain data object classes within the knowledge system are identified by a metadata description derived from a standard schema:



Multiple metadata schemata provide machine usable explanations of a metadata description, which serves to describe the inception and composition of data. Data can come in a variety of flavors, including tabular datasets, videos, images, documents, and other formats. The EDRN Common Data Element (CDE), a standard data dictionary of data elements implemented with the ISO/IEC 11179 standard, provides the elements by which these metadata schemas can be constructed.

The CDEs are vital towards making EDRN applications interoperable. CDEs serve as a standard vocabulary for describing each element including data types, units, representation, enumerated legal values, legal ranges of values, and so forth. With CDEs, software can perform automatic validation, form generation, and scientific correlation.

As an example, consider a typical CDE for "cell count"; a subset of the attributes for this element is as follows:

Attribute	Definition
Contexts	EDRN, validation studies
Object class	Specimen
Property	Bone Marrow
Name	SPECIMEN_BONE_MARROW-CELL_COUNT
Version	1.0
Title	Cell Count
Description	The number of cells (in millions) in a stored cell specimen taken from bone marrow.
Data type	Integer
Minimum value	0
Maximum value	Unlimited
Units	count
Status	Approved

When a piece of data such as "196" is presented with metadata "cell count," software automatically knows how to (1) validate the data when coming from unreliable sources (hand data entry, for example); (2) store the data in short-term storage in an object; (3) persist the data into long-term storage in a database with correct typing; and (4) display the data with appropriate labels and help text.

Every CDE also has its own unique ID in the form of a Uniform Resource Identifier (URI). These take the forms of URLs that point to the CDE's definition pages. Because they're URIs, they can be used in XML-based standards that expect URIs, such as XHTML, XLink, and so forth.

4. Warehousing and Workflow

A fundamental goal of the EDRN Knowledge System has been the capture, processing and warehousing of scientific data generated during validation study. Researchers develop validation studies in an effort to identify biomarkers and disease markers, which may indicate the existence of or potential for cancer. Much of the study centers around the use of sophisticated screening technologies to obtain information from various populations of study. A large set of data samples means more study power in determining correlations in marker and disease activity. As a result, capturing the information for study across the EDRN network is critical.

The EDRN Catalog and Archive Service (eCAS) was initiated as a distributed metadata-driven system for the capture, tracking, processing and retrieval of scientific data from biomarker validation studies. eCAS promises to be an invaluable tool that will make it possible for doctors, scientists, clinicians, and researchers to share their results, correlate their data, discover promising knowledge of new biomarkers, and more; and, as a result, improving the state of public health by combating cancer. As an example, a researcher tabulating results from a spectrograph has access to hundreds of data points. Finding relative minima and maxima creates new data points.

Combining that data with calibration information for the instrument refines that data. Correlating those results with other spectrograph runs creates information for tracking changes over time, between specimens sampled, between instruments manufactured, and more. Each of these states requires a workflow-oriented system for managing the information and the associated processing steps. eCAS implements four specific strategies:

- eCAS tags every datum with an open-ended set of extensible metadata.
- eCAS tracks every datum such that the system knows its location, version, and distribution status at every moment.
- eCAS employs open web standards, including TLS, RDF, XML, and HTTP. eCAS can be connected to science tools for immediate ingestion of data. eCAS can work with departmental data sources without impacting existing operations through translation layers.
- eCAS implements a underlying workflow engine for managing each task.

In the envisioned deployment of eCAS, institutions run eCAS software servers that participate in a peering network grid, creating a virtual organization that transcends institutional boundaries.

This data grid enables scaleable and transparent replication of data and metadata, improving availability and reliability. This architecture is mirrored within institutions, creating virtual departments; and within departments, creating virtual workgroups. Flexible access controls enable researchers to designate that a particular datum be available to specific users, workgroups, departments, organizations, and so forth.

Creating an integrated cancer platform requires the introduction of mechanisms for building pipeline architectures for processing data. We believe this to be a critical capability for NCI as researchers begin to share data sets and one in which progress by the grid community in workflow tools can help. Such a capability will require that researchers be able to reprocess lower level science products in order to verify research results. In addition, EDRN will be able to integrate production of science data products all the way to the instrumentation. This will ensure consistency in how data sets are produced. We will work to make such capabilities available within a grid environment such that remote and server-side processing can occur as part of the integrated platform for EDRN. In addition, high performance and cluster computing for computationally constrained algorithms can be executed remotely on devices that can handle such demands. This will be integrated as part of the capture and distribution mechanism with the eCAS software.

5. Discovery and Access

The component-based architecture of OODT is used in various assemblies to create working applications satisfying specific requirements. In addition to eCAS described above, there are other applications integrated into complete knowledge environment.

EDRN Resource Network Exchange, or ERNE, is a virtual specimen bank unifying disparate and scattered specimen repositories. The **Biomarker Database** is a tracking system for biomarker research, including collection of such data as phase of biomarker development, studies and related trials, and other data. Other applications are already deployed and in development by both JPL and the DMCC.

The eCAS, Biomarker Database, and ERNE applications are all *realized* using combinations of the OODT basic components of catalog/archive service, profile service, product service, and query service. These realizations are made possible using only a thin layer of "glue code" that provides for the communications pathways between the components and visible user interfaces for researchers, doctors, clinicians, and other end users. In addition, by leveraging the common metadata vocabulary (CDEs), these applications are interoperable, supporting discovery of relationships between disparate data and information.

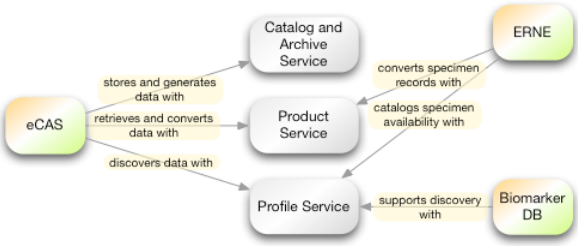
5.1. Component Architecture

The EDRN Catalog and Archive Service, eCAS, is built on the OODT Catalog/Archive Service. It also uses the Profile Service in order to discover additional data. The Profile Service provides eCAS a search function based on metadata that yields matching results in the catalog. The Product Service provides eCAS with the ability to retrieve data from the archive and also performs on-demand conversions.

The EDRN Resource Network Exchange (ERNE) uses OODT Profile Service to optimize its query before being sent to every participating product server. The Profile Service contains metadata that describes what kinds of specimens are available at each ERNE site and various other parameters. For example, when a user searches for blood specimens, ERNE uses the Profile Service to determine which sites have blood. It then queries the Product Service running at each site. At each site, the Product Service's job is to translate the specimen query from the common format into its site-specific format, gather the results, and then translate them into the common format before returning back to ERNE.

While the other EDRN applications make better leverage of OODT's components, the Biomarker Database, being a rather unique application, makes use only of the Profile Service. Its use is in providing a metadata lookup facility in order to attach an open-ended set of metadata to biomarkers being tracked.

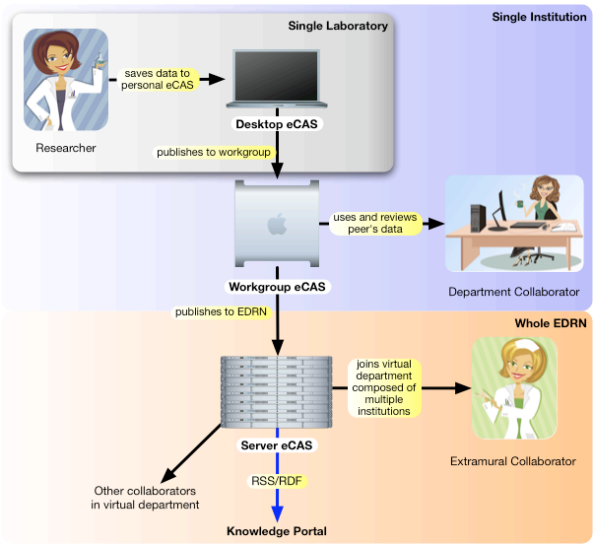
The following diagram depicts the component architecture:



5.2. Application Deployment

Each of the EDRN applications are deployed to participating sites in different ways, reflecting the unique nature of their interactions with users. Where possible, we strive to make installation as simple as possible given constraints on platforms, existing data, and existing processes.

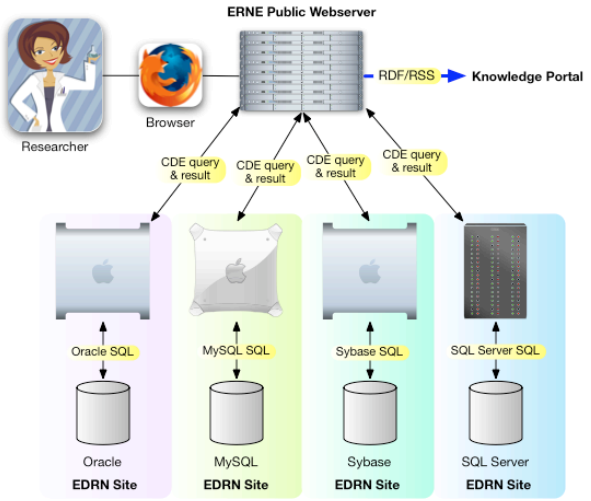
For eCAS, installation may take place on a desktop, workgroup, or server system. Any system running an eCAS installation becomes an eCAS server, capable of accepting data, cataloging it, running post-processing tasks, versioning data, and so forth. The following diagram demonstrates the deployment scenario:



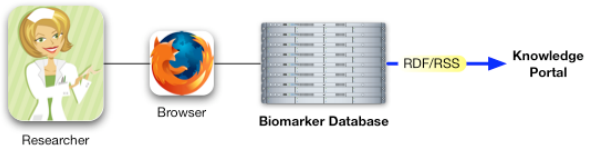
The ERNE application takes a different form from eCAS. Any site that would like to share its specimens first uses a centralized web application to describe how specimens are cataloged and how they relate to a set of CDEs. The CDEs provide a uniform vocabulary for querying for and describing specimens. Using a web application, a site registers the details of its specimen database, including such information as database type and operating system platform, table schema, logical organization of specimens, participants, histology, and other records, and the relationship of such organization to the CDEs.

From this, a software specification is derived that tells a software engineer how to develop a product server that maps an incoming query for specimens in the CDE format to a site-specific query, and how to map site-specific results to CDE results. Sites then install their customized product servers locally with access to specimen databases.

Researchers pose queries for specimens using a centralized web application. That web application then sends the query (in CDE format) to product servers that can answer it. Product servers convert the query, access their local databases, convert the results, and return them to the web application for display. The following figure depicts the deployment:



While eCAS is deployed in a "peer-to-peer" fashion and ERNE is more "client-server," the Biomarker Database is a centralized application. The Biomarker Database enables researchers to track the progress of biomarker development through its various phases. Search functions enable researchers to find interesting biomarkers, participate in clinical trials in support of biomarkers, and share data regarding biomarkers. The following diagram shows the deployment:



Biomarkers are entities tracked within the database whose states change over time based on a specific biomarker development workflow. In this regard, the Biomarker Database is much like an issue tracking system. Attributes of biomarkers are tracked within the database using the CDEs.

Because each of the EDRN applications uses the same set of CDEs, they can interoperate and automatically correlate information. A biomarker tracked in the Biomarker Database can link to a cell

count in a specimen record in ERNE; the result of an analysis can reference that cell count stored in eCAS; etc.

6. Semantic Knowledge Environment

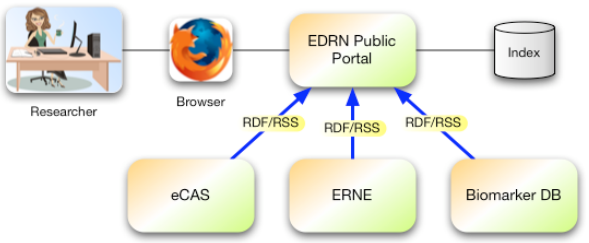
Each of the EDRN applications described thus far—eCAS, ERNE, and the Biomarker Database—all draw from the set of CDEs defined for EDRN. This enables those applications to *interoperate* by sharing a common vocabulary for describing the entities relevant to each application and correlating them with cooperating applications. In addition, these three applications *publish their collective knowledge* to a unified knowledge system portal.

EDRN runs a fourth application in the form of a *public portal* that enables doctors, researchers, clinicians, cancer sufferers, patient advocates, and the general public to examine, browse, search, and track EDRN activities. By publishing the content of the eCAS deployments, of ERNE, and of the Biomarker Database to the public portal, researchers gain a "one stop shopping" location for EDRN data and information with a unified search that presents opportunities for correlation and discovery of cross-application knowledge.

The various eCAS deployments, ERNE, and the Biomarker Database publish their content using the Resource Description Framework (RDF), a semantic web standard. RDF describes content (called "subjects") by first identifying them with URIs. For EDRN, the URIs to eCAS data are the URLs to the eCAS installation plus the path to the data provided by the eCAS server. For ERNE, the URIs are URLs to individual specimen records. For the Biomarker Database, URIs are URLs to single biomarkers tracked in the database.

RDF makes statements about subjects using properties (identified by URIs) and values for those properties. For EDRN, property URIs are URLs to the Common Data Element definitions Disparate resources in eCAS, ERNE, and the Biomarker Database can thus be correlated.

Publication to the EDRN knowledge system portal is planned to happen with a periodic cycle using RSS, a pull-based technology. The portal will request RSS feeds from participating eCAS deployments, from ERNE, and from the Biomarker Database. Using the RDF provided over RSS, the portal will update its internal indexes of what knowledge is available. Researchers posing queries to the portal query those internal indexes. The following diagram demonstrates the architecture:



Queries may be posited to just one type of application's data or to all three. Both forms-based searches and Google-like free-text searches are available. The Google-like free-text search is enhanced with AJAX technology that provides instant feedback on search terms as the user types them, character-by-character. Such feedback is eminently useful in providing hints to the user on the usefulness on search terms without the slow drill-down of search refinement.

7. Related Work

There are several areas work related to EDRN's informatics. The work more or less falls into three fundamental principalities: grid computing, large-scale, data-intensive systems architecture, and bioinformatics. In this section we overview a small cross-section of representative projects in each of the above three areas, comparing and contrasting each project to the EDRN project.

Grid computing is an exciting new paradigm focusing on architecture and the technology for large-scale, parallel computation, and data-intensive operations across loosely connected, large organizations distributed across the world. Grids are divided into two sub-categories. *Computational grids* [2] are highly complex software entities that utilize powerful computing resources and capacious networking capabilities in support of solving extremely challenging scientific tasks. *Data grids* [1], on the other hand, are (also) highly complex software entities, focused conversely on sharing, distributing, and managing large-amounts (terabytes and petabytes) of scientific data.

There are several representative computational grid technologies. Alchemi [3] is a .NET-based enterprise grid computing platform that supports computationally intensive jobs, and service orchestration across distributed compute platforms using web-services infrastructure. JCGrid [4] is a Java-based, open-source computational grid technology hosted at Sourceforge.net. There are also several representative data grid projects. The *Earth System Grid* [5] is a U.S. Department of Energy (DOE) funded project intent on allowing climate modelers at scientific research institutions across the United States to search for and obtain large climate data sets. The EU Data Grid project [6] concluded in 2004 with the goal of

providing middleware infrastructure and technology supporting the sharing and management of petabyte scale data volumes in high throughput grid environments.

Several projects have attempted to define architectures for large-scale, distributed data-intensive systems, including grids. Gomaa et al. [7] present a novel architecture for describing large-scale data-intensive information systems, specifically applied to NASA's EOSDIS science domain. This work focuses only on a single style—federated client-server [8], in contrast to our own work, which has tried to support both federated client-server, and peer-to-peer. Furthermore, Gomaa et al. present no fine-grained mapping of the conceptual architecture to the deployment architecture, or middleware-based solutions for implementing the data-intensive system, again, in contrast to our own work, which has tried to provide detailed architecture to implementation tracing.

The conclusions of two independently conducted research studies [9, 10] identify three key areas that the current grid implementations must address in order to promote data and software interoperability: formality in grid requirements specification, rigorous architectural description, and interoperability between grid solutions. The EDRN project represents a natural leap forward in each of these areas. Its focus on architectures for data-intensive, "grid-like" systems naturally addresses architectural description. Additionally, requirements specification is an integral part of architectural description. Finally, software architecture plays a key role in system interoperability, and the goal of EDRN is to take advantage of the architectural style provided by OODT to ensure interoperability between data-intensive systems constructed using its methodology.

Bioinformatics is involves the use of information technology in furtherance of biology-related scientific tasks. Proteomics studies, specimen tracking, clinical trials all benefit from the support of bioinformatics technologies, e.g., study validation and visualization tools, and patient tracking systems. There are several related bioinformatics efforts to EDRN. Finkelstein et al. [11] describe three key challenges that software engineers must faced in support of bioinformatics and systems biology tasks: (1) defining and managing views of bioinformatics model*s*, (2) providing model checking capabilities and validation, and (3) maintaining consistency amongst distributed information models. In this paper, we have described how *each of these three* key challenges was addressed within the context of the EDRN project. Begent et al. [12] identify six key complementary challenges involved with the large-scale integration of biomedical information systems: (1) *decentralized technology construction*, (2) *unobtrusiveness*, (3) *construction of a flexible system architecture*, (4) *data model and element variety*, (5) *data ownership* and (6) *training of users and engineers*. The EDRN project has been a leader in addressing each of these areas through the use of the OODT software technology. In a recent paper [13], we described how OODT addresses nine key software engineering challenges, including *unobtrusiveness*, *data ownership*, and *flexible system architectures*.

The National Cancer Institute (NCI) has recently begun an initiative known as the Cancer Biomedical Informatics Grid, or caBIG [14]. The goal of the system is to create "a common, extensible informatics platform that integrates diverse data types and supports interoperable analytic tools" to "allow research groups to tap into the rich collection of emerging cancer research data while supporting their individual investigations". Although the domain of caBIG will ultimately include data-intensive systems, there has been no plan to date within that project to establish the corresponding architectural styles appropriate for such systems. Rather, caBIG has leveraged systems and tools already in use (e.g., Globus [15]) in order to create its initial infrastructure rapidly. Thus, the EDRN project is highly complementary to the on-going work of caBIG and, conversely, EDRN project can use the experience from caBIG to better identify the architectural needs of the biomedical community. In particular, our recent work within EDRN has involved the integration of one of caBIG's core software technologies [16] into EDRN.

8. Conclusion

The EDRN project solves multiple problems of data management. It prevents loss of data, encourages collaboration, facilitates discovery, enables novel correlation between datasets, and makes data reuse a tangible possibility. Using flexible access controls, metadata, grid features, and pipeline processing, EDRN promises to become a worthy large-scale tool for scientific research, ultimately aiding in the timely eradication of cancer.

9. References

[1] A. Chervenak, I. Foster, et al., "The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Data Sets," *Journal of Network and Computer Applications*, pp. 2000.

[2] C. Kesselman, I. Foster, et al., "The Anatomy of the Grid: Enabling Scalable Virtual Organizations," *Intl' Journal of Supercomputing Applications*, pp. 1-25, 2001.

[3] A. Luther, R. Buyya, et al., "Alchemi: A .NET-based Enterprise Grid Computing System," in *Proc. of Intl' Conference on*

Internet Computing, Las Vegas, NV, USA, pp. 2005.

[4] "Sourceforget.net: Project Info - Java Grid Computing," 2004.

[5] D. Bernholdt, S. Bharathi, et al., "The Earth system grid: supporting the next generation of climate modeling research," *Proceedings of the IEEE*, vol. 93, pp. 485-495, 2005.

[6] "The DataGrid Project: http://eu-datagrid.web.cern.ch/eu-datagrid/," 2006.

[7] H. Gomaa, D. Menasce, et al., "A Software Architectural Design Method for Large-Scale Distributed Information Systems," *Journal of Distributed Systems Engineering*, pp. 1996.

[8] R. Fielding, "Architectural Styles and the Design of Network-based Software Architectures," Ph.D., University of California, Irvine, 2000.

[9] A. Finkelstein, C. Gryce, et al., "Relating Requirements and Architectures: A Study of Data Grids," *J. of Grid Computing*, vol. 2, pp. 207-222, 2004.

[10] C. Mattmann, N. Medvidovic, et al., "Unlocking the Grid," in *Proc. of 8th ACM SIGSOFT International Symposium on Component-based Software Engineering*, St. Louis, MO, pp. 2005.

[11] A. Finkelstein, J. Hetherington, et al., "Computational Challenges of Systems Biology," *IEEE Computer*, vol. 37, pp. 26-33, 2004.

[12] R. Begent, J. M. Brady, et al., "Challenges of Ultra Large Scale Integration of Biomedical Computing Systems," in *Proc. of 18th IEEE International Symposium on Computer-Based Medical Systems*, Dublin, Ireland, pp. 2005.

[13] C. Mattmann, D. Crichton, et al., "A Software Architecture-Based Framework for Highly Distributed and Data Intensive Scientific Applications," in *Proc. of International Conference on Software Engineering (ICSE)*, Shanghai, China, pp. 2006.

[14] "National Cancer Institute. http://cabig.nci.nih.gov/, February, 2004."

[15] "The Globus Alliance, http://www.globus.org/," 2006.

[16] S. Kelly, "ERNE Interface to caTissue, http://oodt.jpl.nasa.gov/wiki/display/edrn/ERNE+Interface+to+caTissue," 2006.